

Agent Simulation of Peer Review: The PR-1 Model

Francisco Grimaldo¹, Mario Paolucci², Rosaria Conte²

¹ Departament d'Informàtica
Universitat de València
Av. Vicent Andrés Estellés, s/n, Burjassot, Spain, 46100
francisco.grimaldo@uv.es

² Italian National Research Council (CNR)
Institute of Cognitive Sciences and Technologies (ISTC)
Viale Marx 15, Roma, Italy, RM 00137
mario.paolucci@istc.cnr.it, rosaria.conte@istc.cnr.it

Abstract. Peer review lies at the core of current scientific research. It is composed of a set of social norms, practices and processes that connect the abstract scientific method with the society of people that apply the method. As a social construct, peer review should be understood by building theory-informed models and comparing them with data collection. Both these activities are evolving in the era of automated computation and communication: new modeling tools and large bodies of data become available to the interested scholar. In this paper, starting from abstract principles, we develop and present a model of the peer review process. We also propose a working implementation of a subset of the general model, developed with Jason, a framework that implements the Belief-Desire-Intention (BDI) model for multi agent systems. After running a set of simulations, varying the initial distribution of reviewer skill, we compare the aggregates that our simplified model produces with recent findings, showing how for some parameter choice the model can generate data in qualitative agreement with measures.

1 Introduction

Science is both a method - a logically coherent set of norms and processes - and a social activity, in which people and organizations endeavour to apply the method. One of the most important elements of the social structure of science is peer review, the process that scrutinizes scientific contributions before they are made available to the community.

As with any social process, peer review should be the object of scientific investigation, and should be evaluated with respect to a set of parameters. Common sense would suggest, at least, considerations of fairness and efficiency. In addition, two specific dimensions very relevant to research are innovation promotion and fraud detection. Science evolves by revolutions [6], and peer review should be evaluated with respect to its reaction to novelty. Is the current system of peer review supporting radical innovation, or is it impeding it?

Fraud detection, especially for politically relevant matters as medicine and health, is also extremely important; its actual effectiveness at ensuring quality has yet to be fully investigated. In [7], the review process is found to include a strong “lottery” component, independent of editor and referee integrity. While the multiple review approach to a decision between two options is supported by Condorcet’s jury theorem, if we move beyond simple accept/reject decisions by considering scoring and ranking, we find several kinds of potential failures that are not waived by the theorem.

These questions are particularly relevant right now, because, on the one hand, peer review is ready to take advantage of the new information publishing approach created by Web 2.0 and beyond. On the other hand, we perceive a diffuse dissatisfaction of scientists towards the current mechanisms of peer review. This is sometimes testified just anecdotally; list of famous papers that were initially rejected and striking fraudulent cases abound. Leaning on examples is an approach that we do not support because it is known to induce bias [13]. However, some recent papers have shown some numerical evidence on the failures of peer review [4].

In fact, peer review is just a specific case of mutual scoring. Following [9, 10], it is a reciprocal and symmetric type of evaluation which includes narrow access and transparency to the target (at least this is how it is designed in the case of teamwork, see the example of scientific research evaluation). Peer review is the standard that journals and granting agencies use to ensure the scientific quality of their publications and funded projects.

The question that follows is then - can we improve on this process? We are not going to fall for the technology trap, and just suggest that by updating peer review to the Web X.0 filtering, tagging, crowdsourcing, and reputation management practices [10], every problem will disappear - in fact, change could make the problems worse; consider for example the well known averaging effect of searching and crowd filtering [3].

Instead, we propose to create a model (or better, a plurality of models) of peer review, that takes into account recent theoretical developments in recommender systems and reputation theories, and test “in silico” the proposed innovations. In this work, we draw an overview of how we foresee such a model, and we present a first, partial implementation of it. Although the literature of simulation models about peer review is scarce, we highlight the results presented in [12], where the authors focus on an optimizing view of the reviewer for his or her own advantage, and those of [8], aimed at encouraging “good” research behaviour by means of the OpinioNet reputation model.

The rest of the paper is organized as follows: the next section outlines a general model of peer review as well as a restricted model focusing on the roles of the reviewer and the conference. Section 3 explains how the latter has been implemented as a Multi-Agent System (MAS) over Jason [2]. In section 4 we show the aggregates that our simplified model produces when varying the distribution of reviewers’ ability. Finally, in section 5 we state the conclusions of this work and discuss about future lines of research.

2 Description of the proposed model

In this section, we draw the outline of a model of peer review (PR-M in the following) and of its subset that we have implemented. We use agent-based simulation as a modelling technique [1]. With respect to statistical techniques employed for example in [4] or [7], the agent-based or individual-based approach allows us to model the process explicitly. In addition, it helps focusing on agents, their interaction, and possibly also their special roles - consider for example the proposal in [7] of increasing pre-screening of editors or editorial boards. Such a change is based on trust in the fair performance of a few individuals who take up the editors role. Thus, these individuals deserve detailed modeling, that could allow us to reason on their goals and motivations [5].

In this model, we want to catch the whole social process of review, and not just the workings of the single selection process. We will try to simulate the whole lifecycle of peer review, that will allow for example - in the complete model - to reason about role superposition between author and reviewer. This approach distinguishes our effort from that of other authors like [4].

2.1 PR-M

The key entities in our system are: the *paper*, as the basic unit of evaluation; the *author* of the paper; and the *reviewer*, which participate in a program committee of a specific *conference*. We define them in the following paragraphs.

Paper. Here, we do not focus on research but on its evaluation. We assume that the actual value of a paper - that we take as the basic research brick - is difficult to ascertain. While we are aware that this is a strong simplification, we give to each paper a numerical value³, we speculate that the value is only accessible through a procedure that implies the possibility of mistakes.

As a consequence, value is hidden by noise and evaluating papers is modeled as a difficult task - though, noise can obviously be canceled by repeated independent evaluations. In our model, we give papers an intrinsic fixed value. But there is another, different value that can be calculated and that changes in time: the number of citations that the paper receives.

The value of a paper as the number of its citation should, in an ideal case, reflect its actual value. In PR-M, we plan to implement a citation system so that approved papers can be cited by other papers, thus creating a network of citations. The decision process will be carried on by the simulated author. With both an intrinsic value and a citation count, after an initial bootstrapping phase, we could check the correlation between these two. The larger the correlation, the better the whole system of peer review is performing.

³ How to modify this assumption is an interesting research question. We can imagine creating a vector of quality dimensions as presentation, technical quality, and innovation; and also removing the assumption completely - to a point in which we could say that papers do not exist, but only their interpretation. As a first attempt, and to provide a reference, using a single value seems to be the obvious choice.

Author. Authors create papers and submit them to the conferences. With the citation network, the author will also decide on what papers are to be included in the bibliography. We plan to develop a probabilistic choice where a paper will have a higher chance to be cited depending on a list of factors including paper presence in a conference where the author is in the PC, or has submitted a paper; being co-authored by the author himself; and being a highly cited paper, thus mirroring the positive feedback mechanism that operates in research. Authors could have individual preferences on the weights. By varying the distribution of the intrinsic value of the papers submitted as well as the author preferences, the PR-M model will allow us to analyze the evolution of the quality of the papers published by each conference.

Reviewer. Reviewers can be part of the program committee (PC) of any number of conferences. In every simulation cycle, representing one year or conference edition, they evaluate a certain number of papers for each conference that enlists them in the PC.

The PR-M model characterizes reviewers by a probability value, named reviewer skill (s), that represents the chance they actually understand the paper they are reviewing.

The distribution of s values is the primary cause of reviewing noise. We will experiment with several distributions, including a uniform distribution of s values across reviewers (which we consider a low level of noise in evaluations) and other, left-skewed distributions where a low level of reviewing skill is more frequent.

Conference. As with the paper, we use the term conference in a general sense; it covers also, for example, the journal selection process. The authors' decision about what conference to send their works to is crucial, since the number of papers received is a measure of success for the conference, and their quality will determine the conference's quality. Can the review-conference system ensure quality in the face of very strong noise, variable reviewers skill, thanks to some selection process of PC composition that leans on the simplest measurable quantity - disagreement? The number of evaluations a paper receives are just a few - three being a typical case. Thus, the conference is where all the process comes together - are three reviews enough to cancel noise? For what distributions of papers and reviewers skill?

2.2 PR-1

In this paper, we only present a restricted implementation of the full model. This restricted model, that we call PR-1, contains a subset of the features in PR-M, focusing on the roles of the reviewer and the conference only. Thus, the authors and the papers are not included in the following PR-1 definition.

PR-1 represents the peer review problem by a tuple $\langle R, C \rangle$, where R is the set of *reviewers* participating in the PC of a set of *conferences* C .

Each reviewer $r \in R$ has an associated skill value $s \in [0, 1]$. The result of reviewing is accurate with probability s , and completely random with probability $(1 - s)$. To test different distributions in the unit segment, we use the beta distribution. Depending on its two parameters (see figure 1), this distribution can easily express very diverse shapes such as: a uniform skill distribution ($\alpha = 1, \beta = 1$); a set of moderately low skill reviewers ($\alpha = 2, \beta = 4$), and a mix of very good and very bad reviewers ($\alpha = 0.4, \beta = 0.4$).

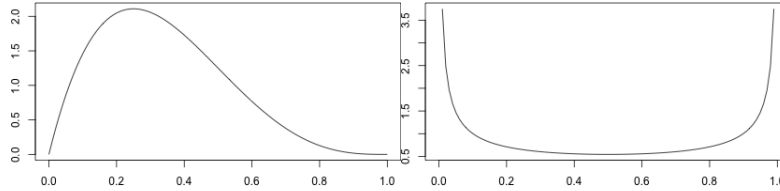


Fig. 1. Beta distributions used in the paper. From left to right, values for (α, β) : $(2, 4)$ corresponding to low skill reviewers, $(0.4, 0.4)$ corresponding to a mix of very good and very bad reviewers. The uniform distribution, also used in the paper, is not shown.

Conferences $c \in C$ are represented by the tuple:

$$c = \langle np, rp, pr, R_c, pa, ac, I, d, e \rangle$$

Each conference receives a number of papers np every year, and employs a subset of reviewers $R_c \subseteq R$ to prepare rp reviews for each paper. The size of the PC ($|R_c|$) depends on the number of reviews done per PC member pr .

Papers have an associated value representing their intrinsic value, and receive a review value from each reviewer. The intrinsic values follow a uniform distribution over a N -values ordered scale, interpretable as the standard from strong reject to strong accept scores. Conferences accept the best pa papers whose average review value is greater than the acceptance value ac . That is, pa determines the size of the conference measured in terms of the number of papers accepted.

We also introduce a PC update mechanism in the face of disagreement. The purpose here is to test a mechanism that could contribute to improvement of papers quality by removing PC members that disagree with others. While there is no such mechanism currently in operation - at least openly, because PC formation is left in most cases to the arbitrary decisions of chairs - in the era of electronic reputation systems, it is not strange to think of ways to make reviewers, in some sense, accountable.

Thus, after the reviewing process, the conference updates the images $i \in I$ of each reviewer in R , according to the disagreement with the other reviewers of the same paper. This disagreement is calculated for each paper as the difference between the review value given by the reviewer and the average review value for that paper. When this difference gets higher than a disagreement threshold d , the reviewer disagreement count grows by one; values are recorder in an image

representation of the form $i = \langle r, nd, nr \rangle$, where r is the reviewer, nd is the accumulated number of disagreements and nr is the total number of reviews carried out. These images are then used to discard the e reviewers with a higher ratio nd/nr and select e new reviewers from R . This way, conferences perform a selection process which selects reviewers who provide similar evaluations. Given our choice for reviewers' mistakes (if they don't understand the paper, the evaluation is random), this mechanism should also select good reviewers.

3 Implementation details

The PR-1 model has been programmed as a MAS over Jason [2], which allows the definition of BDI agents using an extended version of AgentSpeak(L) [11]. This MAS represents both conferences and reviewers as agents interacting in a common environment. Thus, the PR-1 MAS can be configured to run different simulations and evaluate the effects of the parameters in the PR-1 model. For instance, the code in Table 1 shows how to launch a simulation with 10 conferences and a pool of 500 reviewers.

Table 1. The PR-1 MAS launcher configured for 10 conferences and 500 reviewers.

```

1  MAS pr_1
2  {
3      infrastructure: Centralised
4
5      environment: env.PeerReviewEnv(10, 500)
6
7      agents:    conference #10;
8                  reviewer #500;
9
10     aslSourcePath: "asl";
11 }

```

The reviews carried out by the pool of reviewers can be simply programmed in AgentSpeak(L) as shown in Table 2. Here, we use the belief `skill` to set the skill value associated with each reviewer. As already mentioned, we can change the distribution of these values through the (α, β) parameters of a beta distribution (lines 1–3). Each time the reviewer has to review a paper, the `+?review` test plan is executed (lines 6–11). Then, the review is accurate with probability S , and completely random with probability $(1 - S)$.

Conferences can be configured through a set of beliefs in the `conference.asl` file. Table 3 shows the ontology of beliefs used to set parameters such as: the number of papers received (`n_received_papers`), how many of them can be accepted (`max_papers_accepted`) or the number of discordant reviewers that get substituted per year (`n_reviewers_exchanged`). Additionally, a set of image beliefs will be managed by each conference in order to represent the images of the

Table 2. reviewer.asl file defining the reviewer’s behavior.

```

1 skill(tools.beta(1,1)).           // Uniform distribution
2 // skill(tools.beta(2,4)).        // Low skill reviewers
3 // skill(tools.beta(0.4,0.4)).    // Polarized reviewer skill
4
5 // Plan to review papers
6 +?review(IdPaper, Value, Review) : skill(S) & paper_scale_values(N)
7     <- if (math.random < S)
8     {
9         Review = Value
10    } else {
11        Review = math.floor(math.random(N)) + 1
12    }.

```

reviewers in the pool. In addition to the previous beliefs, the conference.asl

Table 3. The ontology by the conference agents.

Belief formula	Description
n.received.papers(NP)	<i>NP</i> is the amount of papers received by the conference.
reviews.x.paper(RP)	<i>RP</i> is the number of reviews done for each paper.
papers.x_reviewer(PR)	<i>PR</i> is the number of papers reviewed by each reviewer.
max.papers.accepted(PA)	<i>PA</i> is the maximum number of papers the conference accepts.
paper.scale.values(N)	<i>N</i> is the scale of values for the papers.
accept.value(AC)	<i>AC</i> is minimum value for a paper to be accepted.
image(R, ND, NR)	<i>R</i> is the number of the reviewer, <i>ND</i> is the accumulated number of disagreements, and <i>NR</i> is the number of reviews done by reviewer <i>R</i> .
disagreement.threshold(D)	<i>D</i> is the disagreement threshold to punish reviewers.
n.reviewers.exchanged(E)	<i>E</i> is the number of reviewers exchanged each year.

file contains the set of plans dealing with the goals involved in the peer review system. Table 4 shows some snippets of the plans in this file. For instance, the plan `!celebrateConference` (lines 1–4) first launches the subgoal related to the reviewing process (`!reviewProcess`). For each paper received, a number of *RxP* reviews are collected (line 11). Then, the conference accepts the best *PA* papers (lines 30–34) amongst those exceeding the acceptance value *AC* (lines 13–17). The image of the reviewers in the PC is updated according to the disagreements with the other reviewers of the same paper (lines 19–27). These new images will be used to satisfy the goal of updating the members of the PC (`!updateReviewers`) in line 3.

4 Results

As a proof of concept, in this paper we show what happens in our simplified model (PR-1) if we change the distribution of reviewers’ abilities. Thus, we

Table 4. Plan snippets from the conference.asl file.

```

1  +!celebrateConference(Year)
2    <- !reviewProcess;
3      !updateReviewers;
4      !!celebrateConference(Year + 1).
5
6  +! reviewProcess : n_received_papers(RP) & reviews_x_paper(RxP) &
7    accept_value(AC) & max_papers_accepted(PA) &
8    disagreement_threshold(D) & ...
9    <- for ( .range(PaperId,1,RP) ) {
10      PaperValue = math.floor(math.random(MaxValue)) + 1;
11      for ( .range(1,1, RxP) ) { /* Ask for reviews ... */ }
12      // Evaluate the paper
13      .findall(Review, review(PaperId,_,Review), Reviews);
14      AvgReview = math.average(Reviews);
15      if ( AvgReview > AC ) {
16        +accepted_paper(PaperId, PaperValue, AvgReview);
17      }
18      // Update the image of the reviewers
19      for ( review(PaperId, R, Review) ) {
20        ?image(R, ND, NR);
21        .abolish(image(R,_,_));
22        if ( math.abs(AvgReview - Review) > D ) {
23          +image(R, ND+1, NR+1);
24        } else {
25          +image(R, ND, NR+1);
26        }
27      }
28    }
29    // Limit the number of accepted papers
30    while ( .count(accepted_paper(_,_,_)) > PA ) {
31      .findall(acc_paper(R, PId), acc_paper(PId,_,R), AcceptedPapers);
32      .min(AcceptedPapers, acc_paper(_,PaperIdMin));
33      .abolish(accepted_paper(PaperIdMin,_,_));
34    }

```

experiment with different initial probability distributions for the only characteristic of reviewers' skill, that is, the probability that a reviewer gets his/her paper right. We consider three cases, that is, uniform ability, low average skill, and polarized skill⁴. The shape of the beta distributions that we apply are shown in Fig. 1.

For this first set of experiments, we have ten conferences (which are essentially the same) receiving 100 submissions each (np), drawn from a uniform distribution. Papers are assigned an intrinsic value in a 10-values ordered scale, interpretable as the standard from strong reject to enthusiastically accept scores. We have fixed $pa = 100$ and $ac = 5.5$, so that all papers whose average review value is greater than 5 are accepted. We have set $rp = pr = 3$, i.e., the same number of reviews per paper and per PC member. Thus, a conference will need as many reviewers as it receives papers. That is, conferences will employ 100 reviewers each from a pool of 500 reviewers. There is no limit to PC memberships

⁴ High average skill is not considered because the uniform distribution already yields a high quality selection process.

for an individual reviewer. Ideally, the same group of 100 reviewers could constitute the PC of all ten conferences. Finally, we use a disagreement threshold of 4 (d) and a 10% reviewer turnover rate ($e = 10$). Substitute reviewers are selected randomly in the pool.

4.1 Measured quantities

For each set of experiments, we measure several quantities, that we present, in their time evolution, in the following figures. The results are presented with five number summary (the central line marks the median, then the successive quartiles), collecting together the data of the different conferences (that are equivalent in PR-1) and in a window of five consecutive years.

The average accepted quality is the primary measure of success for the selection system. Paper quality, if the review process works perfectly, should select 20 top score papers, 20 with quality 9, and 10 of quality 8, leading to an ideal score of 9.2. The worst possible case (papers are accepted completely at random), as a reference value, would simply be the mean of scores from one to ten, amounting to 5.5.

In parallel to the paper selection process, based on disagreement measures between reviewers, program committees are reorganized with the aim to select the best reviewers. Thus, another quantity we measure is the average quality of reviewers that are part of program committees, under different initial conditions for their distribution. In principle, better reviewers should select better papers.

We also show the number of good papers (i.e., with an intrinsic value greater than 5.5) rejected, and the number of bad papers (i.e., with an intrinsic value less than 5.5) accepted. While the previous quantities can be seen as measures of efficiency, these two can be thought of as measures of fairness.

In fact, good papers rejected and bad papers accepted are especially important because of the high-stakes nature of investment that researchers do on each paper - on the one hand, an “out-of-the-blue” rejection can seriously impact career, especially in small research groups; on the other hand, the publication of bogus papers creates a stigma on journals and conferences. Finally, another interesting measure of success for a conference review process had been defined in [4] as *divergence*: the normalized distance between the ordering of the accepted papers, and the ordering induced by another quality measure.

In [4], divergence was calculated with real data of an anonymised “large conference”, comparing review results against paper citation rates registered five years later. We perform a similar calculation, not against citation rates but against our idealized paper quality. The distance used is calculated simply by the (normalized) number of elements ranked in the top (1/3 or 2/3) by the review process that are not in the top (1/3 or 2/3) in the ideal quality ordering. The result for the large conference, that the authors of [4] claim to be disappointingly comparable to random sorting, is a value of 0.63 at 1/3 and 0.32 at 2/3.

Note how this ordering concerns only the set of accepted papers; good rejected papers or bad accepted ones do not enter this calculation. This value can be

considered as another measure of efficiency of the system: the lower it is, the more efficient the peer review.

4.2 Uniform ability

Here we show the results obtained from a reviewer skill distribution with parameters (1,1) - a uniform distribution.

From figure 2, we can see how the quality of accepted papers starts already over the average. The process improves in time for both the paper quality and reviewer skill; however, only the second has a significant effect. The convergence process seems to manage selecting good reviewers, but this happens without a substantial quality improvement. Mistakes in paper evaluations show only a slight decrease. Finally, divergence from the optimal acceptance ordering remains constant - perhaps after a slight improvement in the first years. At about 0.28 and 0.2, it remains far better than the levels 0.63 and 0.32 reported in [4].

4.3 Low average skill

Apparently, our simulated reviewers perform better than the real ones. What if we decrease their average skill, for example drawing them from a beta distribution with parameters (2.0, 4.0), shown in figure 1 (left)? The results are presented in figure 3. With such a bad average reviewer skill, the quality of accepted papers results lower than in the previous case, and the agreement process yields no or little improvement in time - except in reviewers skill, whose increase however does not seem enough to improve the quality of accepted papers. There just aren't enough good reviewers around to make the process work. Good papers rejected and bad papers accepted abound, making up for more than half the body of accepted papers; divergence, ending at 0.6 and 0.25, seems directly comparable to the values in [4].

4.4 Polarized skill

So far we have shown a relatively good selection process, starting with reviewers with uniform distribution, and a relatively bad one, where most reviewers are of low skill. With yet another shape of the skill distribution, we want to measure how effective the agreement process is in selecting good reviewers. To this purpose, we choose an initial distribution with a double peak - in this experiment, as can be seen from figure 1 (right), most reviewers are very bad or very good. We surely have more than enough good ones for a nearly perfect review process - but will the system be able to select them? Figure 4 shows this is indeed the case. This time, the success of the reviewer selection process takes the average paper quality up with it, obtaining better results than in the uniform case. There are nearly no bad papers accepted, nor good papers rejected towards the end. Divergence is similarly affected, leveling at 0.2 and 0.12 at the end.

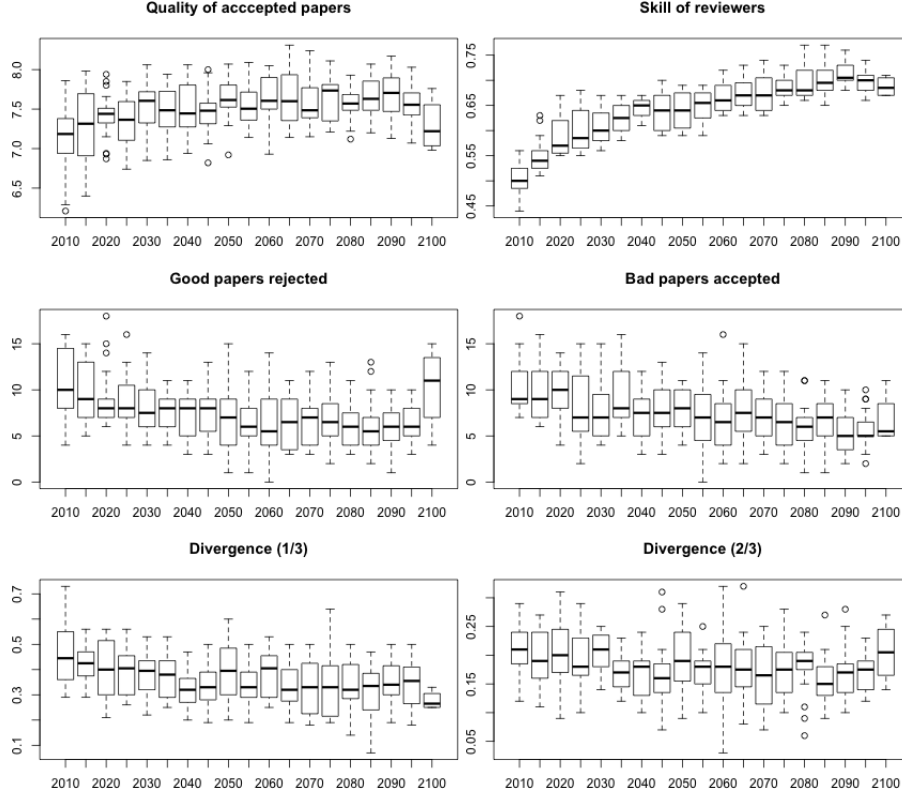


Fig. 2. Results (shown as five-number summary) for a beta distribution with parameters $(1.0, 1.0)$, that is, a uniform distribution, averaged over ten conferences and in periods of five years. *First row*, left, average quality of accepted papers; right, quality of reviewers. Both observable quantities improve substantially in time. *Second row*, left, good papers rejected, right, bad papers accepted, both showing a small improvement in time. *Third row*, divergence values calculated at $1/3$ and $2/3$, both decreasing in time.

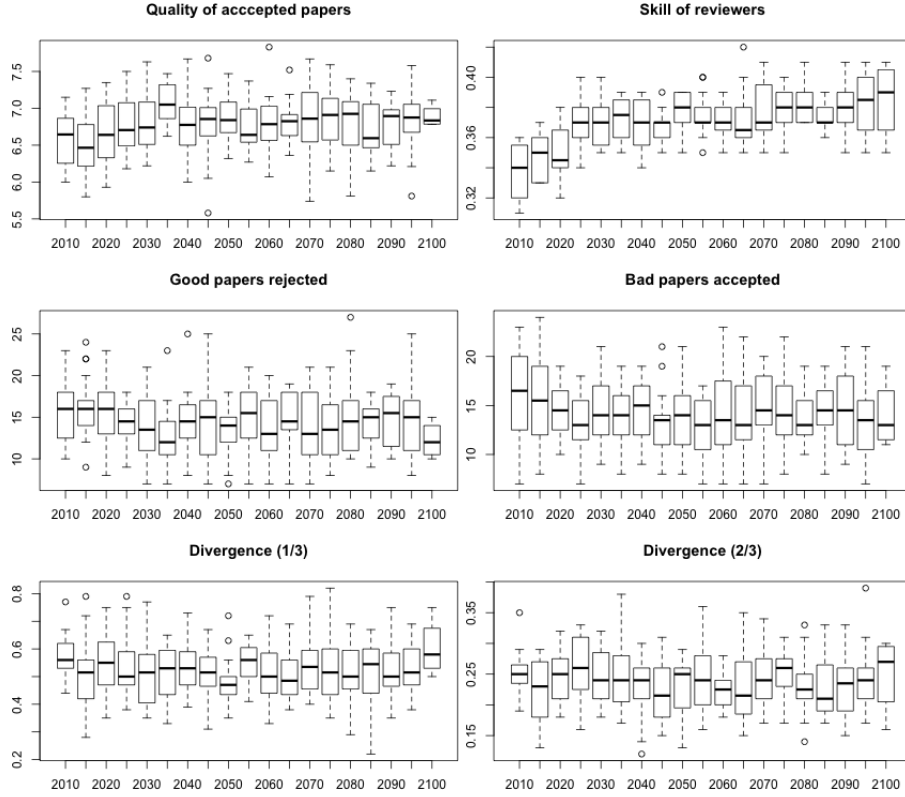


Fig. 3. Results (shown as five-number summary) for a beta distribution with parameters $(2.0, 4.0)$, averaged over ten conferences and in periods of five years. *First row*, left, average quality of accepted papers; right, quality of reviewers. There is no substantial improvement, apart from an increase in reviewers quality. *Second row*, left, good papers rejected, right, bad papers accepted, both stable in time. *Third row*, divergence values calculated at $1/3$ and $2/3$.

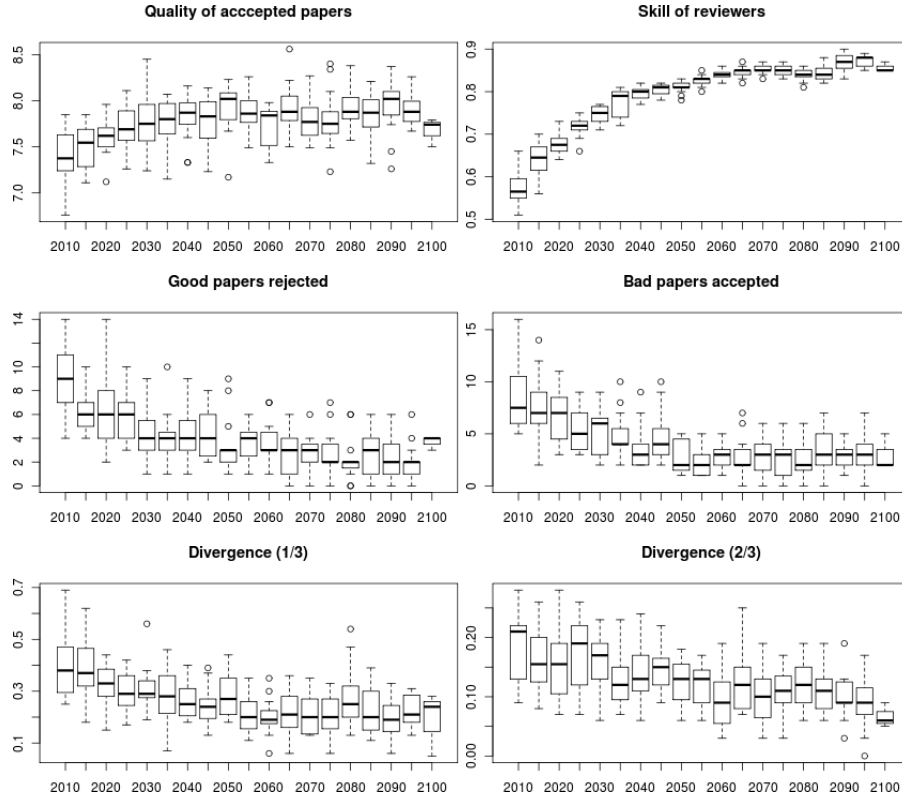


Fig. 4. Results (shown as five-number summary) for a beta distribution with parameters $(0.4, 0.4)$, averaged over ten conferences and in periods of five years. *First row*, left, average quality of accepted papers; right, quality of reviewers. Both observable quantities improve substantially in time. *Second row*, left, good papers rejected, right, bad papers accepted. Both show a marked decrease in time. *Third row*, divergence values calculated at $1/3$ and $2/3$, both decreasing in time.

5 Discussion and Future work

This paper is a first step towards a model of peer review devoted to study and to enhance the way of evaluating scientific research. We have sketched the main elements involved as well as the relations amongst them. A first restricted version of the full model, that we call PR-1, has been implemented as a MAS over Jason. The results show how a conference review process based on disagreement control between reviewers can i) improve in time both the quality of accepted papers and the reviewer skill of PC members; ii) reduce the number of good papers rejected and bad papers accepted; and iii) lower the divergence between the ordering of the accepted papers and an ideal quality ordering. Reviewer selection improves on both the efficiency and the fairness of the review process. The results, for what regards a measure of divergence between reviews and actual quality of the paper, are shown to be qualitatively comparable with the observed data in [4].

Quite a large number of issues still remain open for future work. Limiting PC memberships for individual reviewers and considering role superposition between author and reviewer is one of the next steps. Subsequent versions of the PR model should include the active role of the authors when deciding which conference to send their works to, as it can vary the distribution of the papers submitted to a conference. Furthermore, statistical data coming from conference management systems must be used to validate the results provided by the model.

Acknowledgements

The first author has been jointly supported by the Spanish MEC and the European Commission FEDER funds, under grants Consolider-Ingenio 2010 CSD2006-00046, TIN2009-14475-C04-04 and JC2010-0062.

References

1. E. Bonabeau. Agent-based modeling: methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America*, 99 Suppl 3(Suppl 3):7280–7287, May 2002.
2. R. H. Bordini and J. F. Hübner. Jason. Available at <http://jason.sourceforge.net/>, March 2007.
3. T. Brabazon. The google effect: Googling, blogging, wikis and the flattening of expertise. *Libri*, 56:157–167, 2006.
4. F. Casati, M. Marchese, A. Ragone, and M. Turrini. Is peer review any good? a quantitative analysis of peer review. Technical report, Ingegneria e Scienza dell’Informazione, University of Trento, 2009.
5. R. Conte and C. Castelfranchi. *Cognitive Social Action*. London: UCL Press, 1995.
6. T. S. Kuhn. *The Structure of Scientific Revolutions*. University Of Chicago Press, 3rd edition, December 1996.
7. B. D. Neff and J. D. Olden. Is peer review a game of chance? *BioScience*, 56(4):333–340, April 2006.

8. N. Osman, J. Sabater, and C. Sierra. Simulating research behaviour. In *This volume*.
9. M. Paolucci, T. Balke, R. Conte, T. Eymann, and S. Marmo. Review of internet user-oriented reputation applications and application layer networks. *Social Science Research Network Working Paper Series*, September 2009.
10. M. Paolucci, S. Picascia, and S. Marmo. Electronic reputation systems. In *Handbook of Research on Web 2.0, 3.0, and X.0*, chapter chapter 23, pages 411–429. IGI Global, 2010.
11. A. S. Rao. AgentSpeak(L): BDI agents speak out in a logical computable language. In S. Verlag, editor, *Proc. of MAAMAW'96*, number 1038 in LNAI, pages 42–55, 1996.
12. S. Thurner and R. Hanel. Peer-review in a world with rational scientists: Toward selection of the average. Aug. 2010.
13. J. Wainberg, T. Kida, and J. F. Smith. Stories vs. statistics: The impact of anecdotal data on accounting decision making. *Social Science Research Network Working Paper Series*, March 2010.